



(19)

Europäisches Patentamt
European Patent Office
Office européen des brevets



(11)

EP 0 889 624 A1

(12)

EUROPEAN PATENT APPLICATION

(43) Date of publication:
07.01.1999 Bulletin 1999/01

(51) Int Cl. 6: H04L 29/06, H04L 12/56

(21) Application number: 98305137.6

(22) Date of filing: 29.06.1998

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU
MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: 30.06.1997 US 885072

(71) Applicant: Sun Microsystems, Inc.
Palo Alto, California 94303-4900 (US)

(72) Inventors:
• Hendel, Ariel
Cupertino, California 95014 (US)
• Hejza, Leo A.
Sunnyvale, California 94087 (US)
• Kumar, Sampath H.K.
San Jose, California 95123 (US)

(74) Representative: Harris, Ian Richard et al
D. Young & Co.,
21 New Fetter Lane
London EC4A 1DA (GB)

(54) Trunking ethernet-compatible networks

(57) A network system dynamically controls data flow between physical links by logically combining multiple physical links into a single logical channel trunk, preferably to balance data flow carried by each link. Each link in the trunk has identical physical layer and identical media access control layer characteristics. A system server assigns a single media access control layer address to the single trunked logical channel, preferably randomly by hashing destination media access

control layer addresses for the links. The system server includes, in addition to a physical layer and a network layer, a pseudo-driver software layer disposed therebetween, which pseudo-driver software layer functions as a multiplexer in a receive path and functions as a de-multiplexer in a transmit path. The resultant preferably Ethernet-compatible network system operates in full-duplex mode and distributes packets from the server to the links to preserve temporal order of data flow.

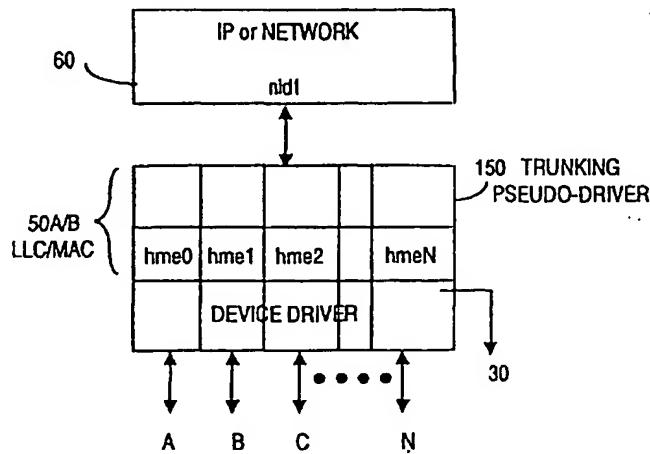


FIGURE 4A

Description**FIELD OF THE INVENTION**

5 This invention relates to networks in general including Ethernet networks, and more specifically to trunked Ethernet and Ethernet-compatible networks.

BACKGROUND OF THE INVENTION

10 A network is a communications facility that permits a number of workstations, computers or other equipment (hereinafter collectively "computer(s)") to communicate with each other. Portions of a network involve hardware and software, for example, the computers or stations (which individually may comprise one or more central processing units, random access and persistent memory), the interface components, the cable or fiber optics used to connect them, as well as software that governs the access to and flow of information over the network. In networks in which data flow is 100
15 Mbits/sec. ("Mbps") or higher, the transmission medium is often fiber optics. In networks in which a slower data rate is acceptable, e.g., 10 Mbps, the transmission medium may be coaxial cable or, as is often the case for an Ethernet network, twisted wires.

In a network, network architecture defines protocols, message formats and other standards to which the computers and other equipment, and software must adhere. Most network architectures have adopted a model comprising functional layers in which a given layer is responsible for performing a specific set of functions, and for providing a specific set of services. Thus, the services provided by each layer and the inter-layer interfaces can define a network architecture. Protocols define the services covered across a layer interface and the rules followed in the processing performed as a part of that service.

Several organizations have proposed models and standards that have been accepted within the networking community. The International Standards Organization ("ISO"), for example, has proposed a seven layer reference model for computer networking that is called the open systems interconnect ("OSI") architecture. Another set of standards has been promulgated by the Institute of Electrical and Electronics Engineers ("IEEE") set of proposed local area network ("LAN") standards known as IEEE Project 802. This model conforms to the seven-layer OSI model, but directly solely to the lowest two OSI layers, namely the physical layer and the data link layer.

30 Figure 1A depicts a network according to the IEEE Project 802 modification to the ISO seven layer model, in which several computers 10, 10', 10" can communicate data to each other over a physical link medium 20, e.g., cable, typically via a data terminating equipment unit ("DTE") 120, 120', 120". Any or all of the DTEs may be a switch, a router, another computer system, etc. In practice, a network may include hundreds or thousands of computers.

In Figure 1A, it is understood that associated with computer 10" (or equivalent) is a similar seven layer ISO model.

35 The bottommost, first or "L1" layer 30 in both the ISO and Project 802 model is a physical layer that is concerned with connections between two machines (e.g., computers 10, 10') to allow transmission of bit streams over a physical transmission medium (e.g., cable 20). Thus, physical layer 30 is concerned with types of cabling, cable plugs, port connectors, and the like. Often a server will have up to fifty physical port connectors, although more modern units such as the Sun Microsystems, Inc. Enterprise model 10,000 server can accept 100 physical links, each having 100 Mbit/sec. flowrate.

40 Many Ethernet networks adhere to a carrier sense multiple access with collision detection ("CSMA/CD") standard. In the 802 model for CSMA/CD, a reconciliation interface 40 defined by a Media Independent Interface ("MII") standard exists for the reconciliation sublayer 40 interface between physical layer 30 and a media access control ("MAC") sublayer 50B.

45 Under MII, data and delimiters are synchronous to the corresponding clock, and two asynchronous media status signals are provided, namely carrier sense ("CRS"), and collision ("COL"). MII provides a two wire serial management interface for control and status gathering, namely management data clock ("MDC"), and management data input/output ("MDIO"). In the OSI seven-layer model, the layer above the physical layer is a data link layer that is responsible for error-free transmission of data frames between network nodes. A data link control protocol describes operation and interfaces of this layer, which must also shield higher layers in the model from concerns about the physical transmission medium.

50 In the 802 model shown in Figure 1A, the data link layer is subdivided into MAC layer 50B and an overlying logical link control ("LLC") layer 50A, collectively a second or L2 layer. The media access control sublayer is concerned with access control methods to determine how to control the use of the physical transmission medium. The LLC sublayer 50A is responsible for medium-independent data link functions and allows the network (or internet protocol, "IP") layer 60 above to access LAN services independently of how the network is implemented. The network or IP layer 60 is often referred to as layer 3, or L3. According to the 802 architecture, LLC sublayer 50A provides services to network 60 in the same fashion as would a conventional data link protocol in a wide area network.

The MAC sublayer 50B provides services to the overlying LLC sublayer 50A, and manages sharing of the transmission medium among the different stations on the network. A media access management function receives a frame from the data encapsulation function after the necessary control information has been added. Thereafter, media access management is responsible for ensuring physical transmission of the data. The data frame in an Ethernet full-duplex environment has a maximum size of 1,518 bytes.

Several 802 standards exist for MAC sublayer 50B, including the so-called carrier sense multiple access with collision detection ("CSMA/CD") standard, and the 802.3 MAC standard provides flow control mechanisms in a half-duplex environment. In such environments, CSMA/CD defines data encapsulation/decapsulation and media access management functions performed by MAC sublayer 50B itself, the data encoding/decoding function being performed by underlying physical layer 30. However, as described later herein, the present invention assumes full duplex operation at the physical and media access control layers 30, 50B for each segment of a trunk.

Physical transmission of the data may be ensured using carrier sensing to defer transmission until the network is clear. In brief, a transmitting station (e.g., computer or user 10) listens or monitors the transmission medium (e.g., cable 20) before transmitting to determine whether another station (e.g., computer or user 10') is currently transmitting a message, e.g., to learn whether the medium is free. Using the services of the L1 physical layer 30, the media access management determines whether the transmission medium (or carrier) is presently being used. If the medium is not being used, media access management passes the data frame to L1 physical layer 30 for transmission. Even after frame transmission has begun, media access management continues to monitor the carrier. If the carrier is busy, media access management continues monitoring until no other stations are transmitting. Media access management then waits a specified random time to allow the network to clear and thereafter begins transmission.

But other station(s) having messages to send may all listen simultaneously, discern that the transmission medium appears quiet, and begin to transmit messages simultaneously. The result is a collision and garbled messages. If signal collision is detected, receiving stations ignore the garbled transmission, transmitting stations stop transmitting messages immediately and transmit a jamming signal over the medium. Following collision, each transmitting station will attempt to re-transmit after waiting for a random backoff-delay time period for the carrier to clear. Thus, a station transmitting must listen sufficiently long to ensure that collision has not occurred.

In Figure 1A, network or IP layer 60 concerns the routing of data from one network node to another. It is the role of network layer 60 to route data between network nodes.

Transport layer 70 provides data transfer between two stations at an agreed upon level of quality once a connection is established between the stations. Transport layer 70 selects the particular class of service to be used, monitors transmission to ensure maintained service quality, and advises the stations (or users) if quality cannot be maintained.

Session layer 80 provides services that organize and synchronize a dialogue occurring between stations, and manages data exchange between stations. As such, session layer 80 controls when stations can send and receive data, based upon whether they can send and receive concurrently or alternately.

Presentation layer 90 ensures that information is presented to network users meaningfully, and may provide character code translation services, data conversion, data compression and expansion services.

Application layer 100 provides a mechanism for application processes to access system interconnection facilities for information exchange. The application layer provides services used to establish and terminate inter-user connections, and to monitor and manage the interconnected systems and the resources they employ.

Although the network shown in Figure 1A may be half-duplex (or shared), or full-duplex, with respect to the present invention full-duplex only will be assumed. In full-duplex, there are transmit and a receive communications paths, and one or more stations may transmit and receive simultaneously. The dual communications channel or path may in fact be multiple wires or cables, or a single wire or cable that simultaneously carries transmit and receive signals in both directions, perhaps using frequency division. Full-duplex networks can provide a higher data rate than half-duplex networks, often 100 Mbps, and are often preferred because of the more rapid communication rate.

Historically, network communication link speeds have increased: 10BASE-T links have given way to 100BASE-T links, which are now being scaled to transmission rates of up to 1 gigabit/second (1000 Mbps) with 1000BASE-T links. But in practice the choice of link speeds (e.g., 10 Mbps, 100Mbps, 1000 Mbps) may not well match the amount of sustained data throughput that a particular network device can support. For example, modern multi-processor servers can sustain greater than 100 Mbps aggregate network transfer rates. Further, when switches and high performance routers are used to interconnect multiple links of a given transmission speed, it is necessary and desired that the inter-switch or inter-router link be able to support at least some aggregation of the links. Enhancing a network link speed from say 100 Mbps to 1000 Mbps is hardly cost effective unless utilization of the higher speed link is enhanced substantially, e.g., perhaps 40% to 50%, or more. Increasing network link speed also requires new hardware.

In Figure 1A, cable media 20 is an Ethernet-compatible cable and will include a plurality of separate wires. As noted, a host, e.g., computer 10, typically will have a plurality of output port connectors (fifty, perhaps), that each accommodate a separate physical cable whose other end will connect to a switch 120 or other DTE. Collectively, the separate wires connecting host 10 to DTE 120 will define a single link of Ethernet-compatible cable, even though

multiple wires are present within the cable.

Figure 1B depicts server 10 with a plurality of output connector ports, here denoted by their network ID numbers, nid0, nid1, nid2, ... nid4. Of course there may be fifty or more of such ports and associated network IDs. The link associated with nid0 is shown physically connected with other equipment 95, the precise equipment and terminating destination being unimportant. A plurality of physical links within cable 20 are shown as terminating at a DTE 120, here a switch. Other ports on switch 120 are shown as having physical links connecting to, for example, yet another system 105, as well as to a client 10', and another DTE 120'.

The operating system associated with server 10 in Figure 1B must treat each of the server connector ports (or the physical cable links connected thereto) on an individual basis. Assume for example that equipment 95 is to be physically relocated from the port position associated with nid0 to some other output port on server 10. To accomplish this, the system administrator must assign a new IP-level address for relocated equipment 95, because cable link A will no longer be used, and a new cable link will be used instead, terminating at a different output port on server 10. Similarly, if cable link B coupling the server output port associated with nid1 to a first input port of switch 120 were down, e.g., broken or defective, no data would be presented to that input port of switch 120 until cable B could be repaired or replaced. There is no automatic re-direction of data to preserve system throughput in the prior art configuration shown.

In Figure 1B, for server 10 to transmit data to client 10', the system administrator would have first had to configure the system such that client 10' is on one of the plurality of sub-links associated with switch 120. In transmitting packet information, server 10 would specify an IP-level sub-link address (associated with IP layer 60) corresponding to one of the links to which client 10' was assigned. As is common, the IP level sub-link address would have two parts: a network ID and a host ID. However each IP-level sub-link address has a different network IP level address associated with it (see also Figure 4B). Understandably, if too many clients seek to use the same IP address, then excess data traffic would attempt to flow through a single link, potentially congesting the network system.

The network administrator will have used a common network level ID for the IP-level address associated with one of the group of links, to which client 10' is dedicated, or bound. One undesired result is that because client 10' DTE and its physical link (here, Q) are bound, the system administrator cannot readily or flexibly group together any arbitrary number of physical links.

The network or IP-level three addresses were denoted nid0, nid1, ... nidN (see also Figure 4B). In the prior art, link grouping is such that sub-links may be denoted at level two with addresses nid1.1, nid1.2, nid1.3, and so forth. To communicate with a piece of equipment, e.g., client 10', that is associated with a group of sub-links including nid1, it would also be necessary to specify which link is involved, e.g., linkage 3, denoted nid1.3.

Thus, software wishing to send information from server 10 to client 10' in Figure 1B must specify the unique destination address as nid1.3. This is because the software application must communicate with level two via level three (see Figure 1A and Figure 4B). In essence, level three is software-commanded to send the data packet to address nid1.3.

Layer three information must be communicated to layer two so as to be understandable to layer two, which understands only its own addressing protocol. Thus, layer three will include a server-maintained mapping function that translates nid1.3 into a physical (e.g., MAC/LLC) address that is meaningful at the network level two layer. In practice, the server causes its layer three to maintain a mapping table (using Address Resolution Protocol or "ARP" information) for each network level ID that is a local link corresponding to an Ethernet ID. In essence, a vertical link level three-to-MAC address level two mapping results for the network ID in question. A separately maintained protocol table will contain an entry for nid1.3 (among other entries) as well as its associated MAC/LLC address. The map-reported particular MAC/LLC address will be and indeed must be used to send the packet in question to client 10'. Unfortunately in a switched network such as has been described, the packets in question have no alternative route to client 10' except this particular MAC/LLC address. For this reason, the addressing system is described as being static.

Figures 2A-2C depict various network interconnect configurations. In Figure 2A, three separate wire link segments A, B, and C are included within media 20, and are used to couple a DTE, here server 10, and a switch 120 (or other mechanism). Of course more or fewer than three wire link segments could be used, depending upon the amount of bandwidth required for the system. Switch 120 is also coupled to several client segments, denoted 1, 2, 3, 4 and 5. In this configuration, server 10 may be replaced with other types of equipment, e.g., a router, a high performance workstation, a printer, etc..

In the embodiment of Figure 2B, three wire link segments A, B, and C connect switch 120' to switch 120, although again a different number of segments could instead be used. In Figure 2C, three segments (A, B, C) couple serves 10' to server 10, although one or both servers could be replaced with other equipment, a router, for example.

In the prior art systems, accommodating configurations of Figures 2A or 2B or 2C typical requires layer 60 to assign to each of the three wires A, B, C, within link 20 a separate IP address. (If multiple clients were to use the same IP address, all of their data traffic would pass on a single link, congesting the system.) Thus, administratively, three internet addresses would be required to handle the three (exemplary) wires within link 20. Network software would then treat each of these links as three different links. In practice, there may be fifty or more separate wires (e.g., A, B, C,)

within a physical link 20, and administratively having to assign and handle separate Internet Protocol addresses for each is burdensome.

For example, in a system having multiple Ethernet links from a server/host system that are connected to an Ethernet-compatible network using an Ethernet switch, the system administrator must statically assign a different ID address (e.g., assign different IP host names) to each link, and also must configure each link with a different MAC address. In addition to being time consuming, the network administrator must also group other host/client systems into different IP sub-nets so that the multiple links can share network data traffic to the server. This grouping of clients into an associated logic group is required because certain clients would have to use one IP address to access the server since one cable is used.

Unfortunately, if a client or DTE had to be relocated within the system, it would now use a new cable coming from the server, and its relocation would require the system administrator to assign a new IP address. The various clients (or other DTE) are distributed at the receiving end of the system, and if the IP address reassignment were not carefully done, too many clients might end up using the same IP address. This could readily result in excessive traffic attempting to pass through a single physical link, congesting the system and degrading system bandwidth. Further, if a physical link were damaged or disconnected, a client whose dedicated IP address involved that link would lose access to the server, there being no other access path. Data intended to go or from such a client would be lost.

Thus, in prior art network configuration protocols, for each physical link the administrator must statically assign a separate level 3 IP layer address and a level two LLC/MAC layer 50A/B address. Because DTE units are statically associated with a link in a group of links, e.g., in a sub-link, an IP layer sub-link address is required.

To summarize, in an Ethernet-compatible network there is a need for a flexible and dynamic method to allow multiple physical links to be logically grouped into a single logical or virtual link. In such a system, the system administrator should be able to assign a single LLC/MAC layer 2 address, and a single IP layer 3 address to the resultant virtual link. Such a system should permit grouping together into a single trunk any arbitrary number of physical links. The resultant system should provide flexibility to the system administrator in terms of rapidly reconfiguring the network system.

Such reconfiguration ability should improve load balancing and the maintenance of network system throughput, and should permit flexible reassignment of data flow traffic to a different physical link in the event there is disruption or failure of data flow through a used physical link. Further, such a system should logically preserve the flow sequence of data packets carried by the network between two communication points. Further, when a previously failed physical link again becomes available, the system should then use such link to help share data flow load over the network trunk.

The present invention provides such a methodology for use in an Ethernet-compatible network system.

SUMMARY OF THE PRESENT INVENTION

Particular and preferred aspects of the invention are set out in the accompanying independent and dependent claims. Features of the dependent claims may be combined with those of the independent claims as appropriate and in combinations other than those explicitly set out in the claims.

In an Ethernet-compatible network, servers and hosts are connected through DTEs, e.g., a switch, using a cable (or equivalent) that carries a plurality of separate wires (or links). Using a rule set, an arbitrary number of the multiple links (or server connectors) are trunked so as to be representable to higher network layers as a single logical (or virtual) link. Such virtual link is assigned one MAC address and one IP address, notwithstanding that a plurality of separate physical links are included and represented. (This is in contrast to prior art techniques in which each connector or link would require a dedicated IP address.)

A virtual link permits system network administration to flexibly and dynamically reassign equipment location without attending to tedious re-addressing tasks. Equipment movement, perhaps because of a link loss, is readily accommodated, and even though a different physical link is used after the move, the moved equipment logically remains within the same virtual network. This flexibility is in contrast to prior art methods in which each separate link in the cable is dedicated and must be configured with its own MAC address and IP address.

An embodiment of the invention logically represents N discrete physical information channels, each bearing a unique physical address, as a single virtual information channel bearing a single physical MAC address. A trunking pseudo-driver software layer is introduced between the IP (or network) layer and the device driver portion of the physical layer, in the Ethernet-compatible network model, solely on the server side. (The client or recipient side is unchanged.) In transmitting packets from a server across the plurality of physical links, load distribution is handled by trunking device driver code within the server. The pseudo-driver code algorithm ensures that transmitted packets are distributed to preserve logical flow order between the two communication end points traversed by the physical links. Physical link failure is accommodated by reassigning traffic to surviving links, and directing at least a portion of packet load to a failed link only when the link is returned to service.

An embodiment of the invention can use the same information that would otherwise be available to layer three

and layer two in deciding what packet address information to send. However, whereas prior art techniques rely upon a vertical mapping link between layer three and layer two, the present invention constructs destination MAC addresses using conventional map-table information, its own source MAC address, and a packet ID that points to layer three. In a level three-to-level two communication, the present invention makes a plurality of level-two links appear to level three to be a single virtual link or device. In contrast to multiple level IP-level three addresses, one per each lower level link, the IP-level three layer in the present invention sees a single virtual or logical device.

An embodiment of the invention can modify host layer two to provide software-implemented drivers having augmented configuration capability. Device driver issue packets with address information that specifies the device corresponding to each link, including trunk-level ID. This information is used locally at the recipient. Because the information 10 packets carried to the recipient are not otherwise different, the recipient is unaware that it is communicating with a trunked link.

An embodiment of the invention assumes that a trunk may have an arbitrary number of segments, each of which has identical full-duplex operating physical layer and media access control layer characteristics, and each of which shares a common source and destination with other trunk segments. Temporal ordering of information packets transported 15 across a given trunk segment is preserved throughout the network, and any broadcast or multicast packets are not echoed or transmitted back. Load balancing is promoted by hashing destination MAC addresses and then selecting an appropriate physical link to which packets will be addressed and sent.

BRIEF DESCRIPTION OF THE DRAWINGS

Exemplary embodiments of the invention are described hereinafter, by way of example only, with reference to the accompanying drawings, in which:

FIGURE 1A depicts a network and model, according to the prior art;

FIGURE 1B depicts typical network system connections, with which the present invention may be practiced;

FIGURES 2A-2C exemplifies various link configurations with which the present invention may be practiced;

FIGURE 3 depicts a modification to the network model of Figure 1A, according to the present invention;

FIGURE 4A depicts the effect of layer modification, according to the present invention;

FIGURE 4B depicts an unmodified layer, according to the prior art.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

An embodiment of the invention simplifies network system administration and promotes system bandwidth without substantially increasing dollar or performance cost by employing new form of trunking. Trunking may be defined as 40 the ability to combine multiple parallel physical links into one logical (or virtual) channel. As used in the present invention, trunks will be limited in that physical links (e.g., media 20) share a common source and a common destination. Further, in such trunks, each of the links (or trunk "segments") are assumed to have substantially identical physical layer 30 and LLC/MAC layer 50A/B characteristics.

Referring again to Figures 2A-2C, physical media 20 includes a plurality of physical links or segments, e.g., wires 45 A, B, C. The present invention operates to make that plurality of physical links or segments appear to be a single logical or virtual link, from the perspective of higher level network layers in the model of Figure 1A. The ability to treat a plurality of physical links as a single virtual link can greatly simplify the task of network system administration, and can also contribute to flexibly increasing and maintaining network bandwidth. Administratively, the network system can use a single IP address and a single Ethernet address for one logical (or virtual) link, even though the link includes a plurality 50 of physically separate cables or segments.

In an embodiment of the invention, within a trunked network operating system, a link is assigned (or is associated with) a single identification (ID) address and a single MAC layer address. One desired result is that a physical cable (or equivalent) comprising a plurality of separate links appears to higher network system levels as one virtual or logical entity, an entity to which IP layer 60 can assign a single IP name.

One advantage of such virtual link referencing is that a higher effective network bandwidth capacity can be flexibly 55 realized. For example, in a network system with ten 10 Mbps links, logically grouping the ten links can produce the throughput of a 100 Mbps link. Of no less importance from the standpoint of a network systems administrator, the addressing benefits provided by the present invention provide great system reconfiguration flexibility.

In implementing an embodiment of the invention, a rule set is defined to which equipment on each end of a trunk must adhere. The rules are symmetric in that each end of the trunk must obey the rules. Further, the rules apply to end stations including DTEs, computers of any classification, etc., and also apply to network infrastructure components (e.g., switches especially).

5 The rule set is as follows:

1. A trunk may have any number of segments, but all segments must have identical physical layer and media access control layer characteristics;
- 10 2. Each trunk segment must share a common source and a common destination with other trunk segments;
3. Temporal ordering of information packets transported across a given trunk segment must be preserved throughout the network, subject only to loss due to bit errors;
- 15 4. Temporal ordering of packets transported across different segments of the trunk can not be assumed;
5. Packets (including broadcast and multicast packets) must not be replicated or duplicated across trunk segments;
- 20 6. Broadcast and multicast packets transmitted through a trunk segment must not be echoed or looped-back to the sender over the other trunk segments;
7. The rules are based upon a model that assumes full-duplex operation at the physical and media access control layers for each segment, and half-duplex operation, using CSMA/CD is neither supported nor desired;
- 25 8. Trunk end stations will associate a single 48 bit IEEE MAC address with all trunk segments;
9. Perfect load balancing across segments can not be assumed, and each trunk end will attempt to load balance across the segments to the best of its ability, subject to all of the foregoing rules.

30 The above rules do not per se address trunk configuration, setup, management, or maintenance, address failure detection or recovery. The embodiment assumes that the physical layer will provide some indication should a particular trunk segment fail. Further, the embodiment of the invention assumes that each trunk end monitors whatever status is provided by the physical layer, and takes appropriate action. In practice, configuration and setup may be performed via manual operations specific to each implementation, although non-manual operations may instead be used.

35 From the server perspective, an embodiment of the invention imposes control over how data flows out, but imposes no control as to how data might arrive. Further, while the server provides a client ID assignment at the device driver level for the logical link, the switch (or other) mechanisms with which the present invention is used must understand the virtual nature of the linkage involved. Thus, the mechanism should realize that although what is seen appears to be an aggregated single link, that the link is a logical link and may involve ten wires that are coupled to ten switch poles. Thus, the switch or other mechanism should know how to distribute traffic from the server to the mechanism preferably evenly among these various links. Typically, switch and other DTE mechanisms are microprocessor controlled and can readily be made to configure a set of specifying mechanism ports as a single logical link.

An important goal in a trunking model is to make even a multi-segment trunk appear to be a single high bandwidth interface from the viewpoint of the server's protocol stack. Further, all clients that communicate with the server via the trunk should have a consistent view of the server's identity, e.g., of the MAC and Internet-Protocol or IP host addresses. Applicants have avoided attempts to achieve load balancing by managing address protocol (ARP) tables within the client. Such an approach is not desirable because the server would have to emit too many ARP frames, and because server ARP implementation would have to be substantially augmented in terms of functionality.

The fourth above-enumerated rule precludes assuming temporal packet ordering in transporting packets across different trunk segments. Accordingly, the server must ensure that all packets in any sequence of packets that requires temporal ordering are transmitted over the same segment of the trunk. In general, transport protocols can recover from situations where packets arrive out of order, but recovery generally entails a significant degradation in throughput. This is because out of order packet reception is treated as an exception, and is not optimized. Accordingly, any server load balancing mechanism should be designed to take advantage of the third rule, namely that temporal ordering of packets transported across a given segment of the trunk must be preserved throughout the network, subject only to loss due to bit errors.

Figure 3 depicts the relatively slight modification to the lower levels of the model shown in Figure 1A for a full-duplex switched environment Ethernet-compatible network, used to implement the present invention. Note that relevant

changes occur below the third model layer, IP or network layer 60. More specifically, a trunking pseudo-driver layer 150 is introduced below IP layer 20 and above LLC/MAC layer 50A/B. Because only slight changes are made to the underlying network model, the resultant network is fully IEEE 802 compatible.

In the trunked server interface of Figure 3, pseudo-driver layer 150 functions as a de-multiplexer in the transmit path, and a multiplexer in the receive path. To satisfy the above enumerated rules regarding temporal ordering, the pseudo driver will attempt to ensure that all of the packets associated with a particular transport layer datagram are enqueued on the same network device transmit queue. This assumption is made on the basis that ordering within a datagram is sufficient, and that ordering between datagrams is unnecessary.

Figure 4A depicts the effect of the modification suggested by Figure 3, and is to be contrasted with Figure 4B, which depicts a conventional, unmodified, IP layer 3, LLC/MAC layer 2 interface. In Figures 4A and 4B, ultimately a plurality of clients (or equivalent), denoted A, B, C, ... N, are served by a server, whose layer 3-layer 2 is shown. However, by slightly modifying the layer two-layer three interface at the server side only, the present invention advantageously allows each client A, B, ... N, to belong to a single subnet, referenced to IP layer 3 via a single IP (or Ethernet) level address, here depicted nid1. In essence, the embodiment of the invention operates peer-to-peer at level two, by including MAC address information on outgoing packets. (By contrast, prior art addressing requires a vertical layer three ARP table-to-layer 2 translation because each physical link is dedicated to a network ID address.)

To the operating system associated with the client of Figure 4A, the N separate system connectors or cables appear to be a single logical link. Administratively, the N pieces of equipment coupled to the server of Figure 4A may be flexibly linked to promote higher system bandwidth. If, for example, there are N 10 Mbit/sec. links, they can be grouped logically to represent an Nx10 Mbit/sec. system. System administration is especially simplified with respect to repositioning pieces of equipment, e.g., moving a DTE from "location" A to location "C" in Figure 4A, or permitting a DTE at location C to receive data even if the physical link that would normally have been used is down.

By contrast, note in prior art Figure 4B, there is a separate IP (or Ethernet) level address for each client, e.g., link nid0 for client A having device hme0, link nid2 for client B having device hme2, and so forth. (See also Figure 1B.) In Figure 4B, the various clients or devices (denoted A, B, ... N) are grouped into N sub-nets with reference to interfacing between level two and level three. Note too in Figure 4B that associated with the IP level three layer and the LLC/MAC level two layer are N network level IDs (nid0, nid1, ... nidN) and N device names, hme0, hme1, ... hmeN, each device name representing a single physical link. As noted, in prior art Figure 4B, IP layer three must recognize a separate network ID address for each physical link.

Consider now the addressing benefits obtained with the present invention. Packets of data incoming to a mechanism (e.g., a switch) carry address information, including a network level 2 destination address and a trunk ID address. Commonly, the destination address is absolutely unique in that no other Ethernet-compatible port in the world will have the same destination address. The destination address is 6 bytes long, the source address is 6 bytes long, and 2 bytes are used to identify the type of packet to be carried by the physical link from server to mechanism. Trunking according to an embodiment of the invention avoids packets having to carry multiply different Ethernet level destination addresses. Thus, in the embodiment of the invention, a switch coupled to receive five input links will not require five separate six byte destination addresses for each of the links.

In addition to assigning a preferably six byte level two destination address, the present embodiment also assigns a level three address. However, whereas the prior art requires one level three internet address for each level two Ethernet address (Figure 4B), the embodiment of the invention functions with a single level three address for a plurality of physical links with a single level two address (Figure 4A). In this fashion, the plurality of level two linkages (or sub-links) are representable as a single virtual link to layer three and higher layers in the network model. The host or server-side layer two is modified to include software device drivers with augmented configuration capability. The modified layer two issues commands to a pseudo-device driver layer, specifying a device corresponding to each link. The relevant trunk is software identified with a trunk ID.

Referring again to prior art Figure 4B, each of the N sub-links (e.g., A, B, ... N) will have at level three a different sub-link IP address. If there were N=5 links, then five different level three sub-link IP addresses would be required. If a host server 10 wished to send data to a client 10' (or other equipment), server 10 would be required to specify an appropriate IP sub-link address that corresponded to the specific link to which client 10' were assigned. Thus, for server 10 to specify client 10', server 10 at IP level three would specify a unique IP address having host ID and network ID. As noted, when the network administrator configured the network, the network level ID assigned for client 10' will be the same as the IP level address that is associated with client 10', because the associated physical link is bound to this client. By contrast, as shown in Figure 4A, a single virtual link spanning layers three and two can represent a plurality of lower level links.

In practice, it would be difficult for the software code implementing the pseudo driver to inspect headers of each packet and attempt to associate them with a particular datagram. However, given a sufficiently fast processor to execute the code, this could be accomplished. It appears sufficient to maintain a small cache of MAC (or IP) destination addresses associated with each network interface, e.g., with each segment of a trunk. Such cache may be maintained

in any of stations 10, 10', 10'', etc. (see Figure 1A), as would be the software code and processor executing such code.

Thus, when the IP layer hands the pseudo driver a packet, the pseudo driver inspects the cache to learn whether it has recently transmitted a packet to the relevant destination address ("DA"). If yes, the pseudo driver will enqueue the packet on the same interface used to enqueue the last packet to this DA. If there has been no recent transmission to this DA, the pseudo driver can either enqueue the packet on the most empty (e.g., least busy) transmit queue, or on the next available queue, e.g., in a round robin fashion. Regardless of the queue it selects, the pseudo driver must update the cache for that queue with the new DA.

Consider now the degenerate case in which a server talks to one and only one client. The present invention still ensures that all packets for that client traverse over the same interface, and hence over the same segment of the trunk. Load balancing is achieved in that the present invention works even better in a non-degenerate case. An ordered delivery of packets is ensured, providing one can assume the degree of interleaving of packets to different DAs between the IP and the network driver is of the same order as the number of processors within a given multi-processor server. As a guide, the cache depth should approximate twice the number of processors within a given server. Of course, as cache depth is increased, cache updating can be somewhat more casual. There may in fact be an optimal cache depth value, as well as an optimal distribution of amount of layer 2 and layer 3 addresses to be cached.

Several switch load balancing mechanisms may be used to forward packets into a trunk. Consider now load balancing among network DTEs such as switches, for which the following set of load balancing guidelines is applicable for switch-to-server trunks (e.g., Figure 2A), and for switch-to-switch trunks (e.g., Figure 2B). These guidelines are intended to ensure that switch behavior is consistent with conventional bridging guidelines. The guidelines are as follows:

1. No frame mis-ordering should occur for a given priority level between a given MAC source and destination;
2. No frame duplication should occur;
3. Load balancing should be transparent to protocols operating above the MAC layer.

One load balancing approach is to emulate a faster link by keeping all trunk segments equally busy. For example, the corresponding output queue could be used as the metric for the busyness of the segment. Providing the links implement flow control, output queue length is a good end-to-end proxy for the segment utilization. (Without flow control, a high segment load would not necessarily be reflected in the state of the output queue, due to packet loss at the other end on the receive queue.)

However using queue length solely to decide what segment to use on a per packet basis could result in frame mis-ordering, in contravention of the first above-enumerated guideline. However, the first guideline will always be satisfied if the segment decision is made solely as a function of source address of the packet, or of the packet's port of arrival. This approach, however, results in a static load balancing function in which trunk effectiveness depends upon distribution of the traffic sources. On one hand, a large number of traffic sources should promote a result having acceptably even distributions. But on the other hand, one might still end up with configurations whose mapping function forwards a substantial amount of traffic to the same segment, which would promote system congestion.

A preferred approach is to implement a dynamic mapping function, while still maintaining frame ordering. This approach assumes that function changes occur more slowly than output queue transit times. By way of example, mapping for a given source address can be determined at the time the first packet with the source address is seen. The mapping can eventually be aged after the source address goes unseen for a sufficient period of time.

Preferably the mapping function considers both the source address and the port of arrival, to reduce the number of pathological cases. For example if traffic is spatially dominated by a particular input port, considering the source address will help spread its traffic. Conversely, considering the port of arrival helps distribute traffic dominated by a small number of addresses (e.g., servers or routers), especially if more than one trunk exists in the switch.

Frame duplication is prevented by treating the set of trunked ports as though they were a single port, with a separate queue per segment, and by ensuring that all forwarding is to only one of the queues. Furthermore, trunked ports are also treated as if they were a single port for purposes of other IEEE 802.1(d) functions such as learning MAC addresses, filtering frames, and executing any applicable spanning tree protocol.

Thus far, the above description has concentrated upon load balancing using MAC layer information. It is also possible for a network switch to observe higher level protocol information to make better load balancing decisions, providing the third guideline (re protocol transparency) is followed. Transparency implies that protocols are not aware of and do not explicitly cooperate with the switch load balancing function. In addition, connectivity must still be guaranteed for protocols that are neither supported nor understood by the switch.

It is practice to implement load balancing based on higher level information for switches that examine Layer 3 headers on a packet-by-packet basis. For example, in VLAN configurations, switches can examine layer 3 headers once and use the corresponding layer 2 information for packet processing. However, applicants have not fully considered potential load balancing performance of such an approach.

Although the foregoing descriptions have been directed to providing a mapping function that promotes equal balancing among segments, the mapping function could be directed otherwise. For example, the mapping function could separate traffic according to priority, or according to whether the traffic is bandwidth managed, or according to best effort. A priority based approach is supported by the first enumerated guideline (no frame mis-ordering) because it is not necessary to preserve packet order across different priorities. A priority based approach is especially straightforward when priority information is well defined at the MAC level, as is the case if VLAN tags are used.

From the foregoing description it is seen that the behavior of a switch or other equipment may be conceptually guided by a set of simple rules. While the particular switch architecture should be taken into account, implementing definitions can be accomplished independently of specific server load balancing behavior.

Consider now implementation of the present invention, within the layer two level on the server-side of an Ethernet-compatible network. In the following description, "hme" nomenclature shall refer to device drivers such as are included in the trunking pseudo-driver layer 150, indicated in Figure 3 and in Figure 4A. In the preferred embodiment, an hme device provides 10/100 Mbps Ethernet-compatible interface on several hardware platforms, including many systems under development by Sun Microsystems, Inc. The hme device on these various platforms is preferably controlled by the same hme device driver binary. A description of an hme device driver and trunking will now be given, as well as a description of such architectural and hme driver software changes as are needed to trunk together multiple hme devices.

By way of overview, the hme device driver is preferably a 2.x Solaris STREAMS-based driver that provides a Data Link Provider Interface ("DLPI") interface. In the preferred embodiment, this driver controls the network interface based on the Fast Ethernet Parallel Port SCSI ("FEPS") ASIC for Sun Microsystems Sbus type platforms, and based upon the Sun Microsystems Cheerio class ASIC for PCI platforms to ensure MAC functionality and Media Independent Interface ("MII") transceivers. The driver also provides ioctl function to set/get driver configuration parameters using the ndd command. The driver source code consists of the following files:

usr/src/uts/sun/io/hme.c - contains all the C functions;

usr/src/uts/sun/sys/hme.h - contains the data structured definitions for the driver;

usr/src/uts/sun/sys/hme_mac.h - contains the definitions of the MAC (FEPS or Cheerio) ASIC

usr/src/uts/sun/sys/hme_phy.h - contains the definitions for the MII transceiver.

Those skilled in the art will appreciate that data structures and functions in the hme driver may be implemented in many ways. What follows is a description of a preferred implementation for some of these structures and functions.

30

35

40

45

50

55

Linked list of STREAMS structures

```

 5      /*
5   * Per- Stream instance state information.
6   *
10    * Each instance is dynamically allocated at open () and
11      *free'd at close ( ) . Each per - Stream
12    instance points
13
15    *to at most one per-device structure using the sb_hmep
16    *field. All instances are threaded together into one
17    list
18
20    :--*of active instances ordered on minor device number.
21    */
22
23    struct hmestr {
24
25        struct hmestr *sb_nextp; /* next in list */
26        queue_t *sb_rq; /* pointer to our read queue */
27
28        struct hme *sb_hmep; /* attached device */
29        u_long sb_state; /* current DL state */
30        u_long sb_sap; /* bounds sap */
31
32        u_long sb_flags; /* misc. flags */
33        u_int sb_mccount; /* # enabled multicast addrs */
34
35        struct ether_addr *sb_mctab; /* table of multi-
36        cast addrs */
37
38        u_long sb_minor; /* minor device number */
39
40        kmutex_t sb_lock; /* protect this structure */
41    } ;
42
43
44
45
46
47
48
49
50    /*
51     * Linked list of active (in use) driver streams.
52     */
53
54
55    static struct hmestr *hmestrup=NULL;

```

Linked list of hme structures:

```

5      /*
 * HMEDeviceChannelinstancestateinformation.
 * Eachinstanceisdynamicallyallocatedonfirstattach.
10
11      */
12
13      struct hme{
14          struct hme *hme_nextp; /*nextinalinkedlist*/
15          dev_info_t  *hme_dip; /*associateddev_info*/
16          /*data-
17          structurestocontrolthetransceiver,MACandtostorestats*/
20
21      } ;
22
23      -----
24
25      -----
26
27      }
28
29
30
31
32
33
34
35      staticstructhme*hmeup=NULL;

```

All hme instance data structures are linked together using the hme nextp field. The pointer to the head of this list is:

Driver Functions:

40 The driver functions and configurations will now be given.

Configuration:

The functions to configure the driver in the Solaris 2.x kernel are preferably:

45 hmeinfo

- returns info about the device;

50 hmeidentify

- identifies the name of the device as "hme" or "SUNW,hme";

hmeattach
55

- allocates the hme structure for the device instance and initializes with the configuration parameters; also registers the hme interrupt handler function for the device.

5 **hmeddetach**

- de-allocates the hme structure and other memory allocated to the device instance; also unregisters the interrupt handler for the device.

5

STREAMS operations:

The functions to implement the STREAMS interface operations are preferably as follows: Interface:

10 **hmeopen**

- opens a STREAM by allocating a hmestr structure; puts the STREAM in DL_UNATTACHED state;

15 **hmeclose**

- closes a STREAM by de-allocating the hmestr structure and all other memory for that STREAM;

20 **hmewput**

- called by the upper-level module to pass a STREAMS message that may be a DLPI primitive or M_DATA message for fast-path operation or a special STREAMS message; M DATA messages are immediately transmitted by calling the hmestart() function; DLPI primitives are enqueued in the hme write queue to be processed by the service function hmewsrv() when the write queue is enabled for servicing;

25 **hmewsrv**

- called by the STREAMS scheduler to process the messages enqueued in the hme write queue; this queue contains M_PROTO type messages and transmit messages waiting for free entries in the transmit descriptor ring; it calls hmestart() to transmit packets, and hmeproto() to process the M_PROTO type messages;

30 **hrmeioctl**

- called by the upper-level modules to process the raw-mode, fast-path, and ndd set and get ioctls.

35 **DLPI Operations:**

The following functions are used to handle DLPI primitives:

40 **hmeproto**

- calls the proper routines to process the DLPI primitive message; calls the hmedreq() function to process DL_UNITDATA_REQ type messages that contain data to be transmitted;

45 **hmeared**

- processes the DL_ATTACH_REQ primitive, which initializes the sb_hmep field of the hmestr structure for the STREAM to point to the hme structure of the specified hme device instance; also puts the STREAM in DL_ATTACHED state;

50 **hmedreq**

- processes the DL_DETACH_REQ primitive; detaches the STREAM from the associated hme device and puts the STREAM in DL_UNATTACHED state;

55 **hmebreq**

- processes the DL_BIND_REQ primitive; associates the STREAM with the specified DLSAP (Data Link Service Access Point) value, which is the type field in the Ethernet header; also puts the STREAM in DL_IDLE state

and enables packet transmission and reception;

hmeubreq

- 5 - processes the DL_UNBIND_REQ primitive; removes the DLSAP associated with the STREAM and sets the STREAM state to DL_UNBIND_REQ;

hmeireq

- 10 - processes the DL_INFO_REQ primitive; returns device information such as device type, ethernet address, broadcast address, etc.;

hmeponreq

- 15 - processes the DL_PROMISCON_REQ primitive to enable the STREAM to receive either all packets (DL_PROMISC_PHYS mode) or all multicast packets (DL_PROMISC_MULTI mode) or all SAP values (DL_PROMISC_MULTI mode);

hmeoffreq

- 20 - processes the DL_PROMISCOFF_REQ primitive to turn-off the specified promiscuous modes;

hmeermreq

- 25 - processes the DL_ENABMULTI_REQ primitive to enable the reception of the specified multicast address;

hmedmreq

- 30 - processes the DL_DISABMULTI_REQ primitive to disable reception of the specified multicast address;

hmepareq

- 35 - processes the DL_PHYS_ADDR_REQ primitive to return either the factory address (DL_FACT_PHYS_ADDR requested) or the current address (DL_CURR_PHYS_ADDR requested);

hmespareq

- processes the DL_SET_PHYS_ADDR_REQ primitive to set the specified ethernet address;

hmeudreq

- processes the DL_UNITDATA_REQ primitive to package data in the message into an Ethernet packet and calls the hmesstart() function to transmit the packet.

45 Hardware Initialization:

hmeinit

- 50 - initializes the hardware (MAC and the transceiver) and initializes software resources for transmitting and receiving packets; is called first time when the device is attached to a STREAM, and is also called when any hardware reprogramming is required, for example, to enable/disable promiscuous mode, to enable/disable multicast reception, to change link speed and mode etc.; also called from the interrupt handler to recover from fatal errors;

55 hmouninit

- called to disable the hardware from receiving and transmitting packets, and to disable interrupts;

hme_reset_transceiver

- called from the hmeinit() function to reset the transceiver;

5 hme_try_speed

- called from the hmeinit() function to bring up the link using auto-negotiation;

10 hme_force_speed

- called from the hmeinit() function to bring up the link in the specified speed and mode (forced speed and mode);

hme_check_link

- called using the timeout() interface every 30 seconds to check the transceiver; the hme device driver supports products that use an onboard transceiver connected to a RJ45 connector or an external transceiver connected to an MII (Media Independent Interface) connector; this function polls for change of transceiver and if transceiver has changed, it calls the hmeinit() function to re-initialize the MAC and the transceiver;

20 Packet Transmission and Packet Reception:**hme_start**

- called by the hmewput() for M_DATA type messages (fast-path) and hmeudreq() for M_PROTO messages with DL_UNITDATA primitive, to transmit an Ethernet packet; also calls the hmereclaim() function to release transmit resources if more than some number (e.g., >64) of descriptors have to be reclaimed; if there are no free transmit descriptors, the message is enqueued at the top of the STREAMS write queue and sets a flag for the queue to be enabled and scheduled for processing again when free transmit descriptors are available;

30 hmeintr

- this routine reads the FEPS/Cheerio status register and checks for events such as errors, packet reception and transmission; normally transmit interrupts are disabled; in case of fatal errors, hme_fatal_error() function is called to update status and to call hmeinit() function to re-initialize hardware and software data structures;
- in the case of non-fatal errors, it calls hme_non_fatal_err() function;
- if packet transmission event is reported, it calls the hmereclaim() function to free the transmit descriptors and the transmit packet buffers;
- if packet reception event is reported, it calls the hmeread() function in a loop to process each received packet;

40 hmereclaim

- this function loops through the transmit descriptor ring from the last reclaimed descriptor until the last packet transmitted; for each transmitted descriptor, it frees the transmitted packet buffer and updates transmit packet count for each packet transmitted; if any packets in the write queue are waiting for free transmit descriptor entries, the write STREAMS queues are enabled for processing by the STREAMS scheduler;

hmeread

- a free STREAMS message buffer is allocated and posted to the hardware in the receive descriptor ring in place of the received packet; if there is no free buffer, the received data are discarded and the buffer is re-posted to the hardware in the receive descriptor ring;
- typically only one stream will be open from the IP layer with sb_sap field value set to ETHERTYPE_IP; the hme structure contains a field hme_ipq to point to the read queue of such a stream; if the type field of the received packet is ETHERTYPE_IP, then this packet is sent to that queue for IP processing;
- otherwise the received packet type field is compared with the sb_sap field of all the streams attached to the hme device and a copy of the packet is sent up all the streams that are eligible to receive the packet.

As noted, implementation of the present invention in a trunked Ethernet protocol requires adherence to the nine

enumerated rules enumerated earlier herein.

In one implementation of the present invention, trunking uses a STREAMS multiplexing pseudo-driver, e.g., software only, and plumbs hme devices that are members of the trunk below it. However, this approach requires some changes in the hme driver and entails the overhead of using a STREAMS putnext() function for communication between the pseudo-driver and the hme driver.

A more preferred embodiment modifies the hme driver itself to support trunk configuration using the ND_SET and ND_GET ioctl interface and the ndd command. This embodiment is preferred as it eliminates much code duplication in the pseudo-driver and avoids the overhead of an additional STREAMS layer.

System booting over a trunk is not possible because the boot code in the fcode does not handle a trunk. The code may transmit boot requests over one of the links, but when the switch sends the responses over multiple links, it does not handle all those packets.

In configuring the trunk, one of the hme devices is identified as the trunk-head and others are identified as trunk-members. For example, consider four hme devices denoted hme 1, hme2, hme3, and hme4, which are to be configured as a trunk. Each of these devices may have different local MAC addresses but to be part of the same trunk all of these devices should be configured with the same MAC address. This is readily achieved by selecting one of the devices as the trunk-head and others as trunk-members. The MAC address of the trunk-head can be selected as the MAC address for all other trunk-members. Further, the instance number for the trunk-head will be used by the user programs and the higher-level protocol modules to access the trunk for data transfers, and other trunk-members will not be accessible for any DLPI protocol operations. For example, if hme 1 is selected as the trunk-head, then the user would create /etc/hostname.hme1 file, so that hme 1 interface is plumbed for TCP/IP protocol access during system boot process.

When a system boots (either from the local disk or over the network using a non-trunked network interface), the kernel configures all network interfaces in the usual fashion as individual links, with local MAC addresses, if any. After kernel configuration is complete, the /sbin/init process executes the /sbin/rcS script, which executes the /etc/rcS.d/S30rootusr.sh script file for configuring network interfaces for which the /etc/hostname.xx files exist. By placing the ndd commands to configure the trunk in a shell script file in the /etc/rcS.d directory and naming the file, by way of example, S20trunk.sh, the init process can be forced to configure the trunk before executing any network interface configuration commands.

Once a trunk is configured, configuration (trunk members or trunk-head) can be modified preferably only after modifying the /etc/rcS.d/S20trunk.sh file and rebooting the system. This restriction helps avoid possible confusion for applications that may maintain a history of network interfaces in the system. Most preferably, a link should be added to a trunk or removed from a trunk after a corresponding reconfiguration operation with the switch to which the trunk is connected.

Configuration Commands:

The hme driver preferably supports the following user-configurable ndd variables for each interface:

- trunk-id - each trunk is identified with a number in the range 1 - 256; 0 value indicates that the interface does not belong to a trunk;
- 40 trunk-head - value of 1 indicates that the interface is the trunk-head; otherwise the value will be 0; this parameter valid only if trunk-id parameter is in the range 1 - 256.

The following sequence of commands in the /etc/rcS.d/S20trunk.sh file illustrate configuration of a trunk consisting of hme1, hme2, hme3, and hme4 interfaces, with trunk-id as 10 and with hme2 as the trunk-head:

```
45      ndd -set /dev/hme 1:trunk-id 10
        ndd -set /dev/hme 2:trunk-id 10
        ndd -set /dev/hme 3:trunk-id 10
        ndd -set /dev/hme 4:trunk-id 10
        ndd -set /dev/hme 2:trunk-head 1
```

50 First all trunk members should be identified with a trunk-id, and then the last command should identify the trunk-head. When the trunk-head is identified, the trunk configuration is completed by re-initializing all the trunk member interfaces using the MAC id of the trunk-head interface. After this step, only hme2 (the trunk-head) will be accessible to the users directly, e.g., only hme2 can be attached to any streams. The other trunk-member -interfaces hme1, hme3 and hme4 cannot be accessed by the users, e.g., cannot be attached to any streams.

55 Once the trunk is configured, the trunk-head hme2 interface is used just like any non-trunked network interface. For example, it can be referenced by the ifconfig command to plumb the TCP/IP protocol stack. All packets sent to the hme2 device by user programs or the IP layer will be sent over one of the links in the trunk such that the order of packets in a flow is maintained. All the packets received on any of the links in the trunk are sent up the streams attached

to the hme2 device.

Only the trunk-head device can be accessed by ndd to change link speed and mode. Even though it is possible to operate the trunk at 10/100 Mbps speeds and in half-duplex/full-duplex modes, it appears best to support only 100Mbps speed and only full-duplex mode.

- 5 Accordingly, in the preferred embodiment, any trunk-member operable only at 10 Mbps speed or in half-duplex mode is treated as if that link were down.

Transmission Load Balancing:

- 10 Messages sent down streams attached to the trunk-head device for transmission should be sent over one of the links in the trunk such that packet flow order in any flow is not broken. Preferably, no more than one-link bandwidth of data should be sent to any particular destination MAC address, unless that particular destination is also known to be trunked. Such will be the case, for example, when the trunk connects two servers directly. Thus, distribution of packets to links will be normally based on the destination MAC address in the Ethernet header.

- 15 The user can indicate when a trunk connects two servers directly by setting a configuration variable in the driver. In such case, packet distribution will be based on the TCP/UDP port number for TCP/UDP packets. Other IP or non-IP packets will be distributed based on the destination MAC address, which advantageously provides for the possibility of distributing different multi-cast streams to different links.

- 20 But as noted in rule number 4 earlier herein, it is important to preserve temporal ordering of packets. A preferred method to preserve temporal order when transmitting packets is to always use the same link, determined by hashing the flow-id (Destination MAC address or the TCP/UDP port number) to the link-id. Many hashing schemes may be used, the goal being to provide a randomizing function.

- 25 A linkmap table is used to map the flow-id to a particular link. The number of linkmap table entries will equal the number of trunk-members (including the link-head) if the number of trunk-members is a power of 2. If not a power of 2, the number of entries should be equal to the next higher power of 2. For example, if there are five trunk-members, there will be eight entries in the linkmap table.

- 30 Each linkmap table entry contains a pointer to a trunk-member and contains some state information for using that trunk-member. If the size of the linkmap table is 2^n , then the last n bits of the flow-id of a Tx packet may be used to index the linkmap table and to select a trunk-member for transmitting the packet.

- 35 At the time of trunk configuration (when the trunk-head is identified), the linkmap table is created and initialized with pointers to trunk-members, preferably in a round-robin fashion. If the trunk size is not a power of 2, some trunk-members are assigned to two entries.

- When a trunk-member link goes down, the entries in the linkmap table that contained pointers to that trunk-member are changed to point to another active trunk-member. The state information for these entries is modified to wait for a 40 timeout before using this new trunk-member for transmission of packets with flow-ids that map to these entries. Such packets are held back in the STREAMS queues until the timeout occurs. Using such timeout helps maintain the temporal order of packets sent out before the trunk-member failure. In contrast to prior art approaches wherein a failed link implies data loss, the present invention can maintain data flow, albeit via a different flow path.

- Similarly, when a failed trunk-member link recovers, at least one entry in the linkmap table will be modified to 45 repoint to this recovered trunk-member. The state for those entries is modified to wait for a timeout before using those entries for further transmissions.

Preferred Data Structures for Trunking:

- 45 The following user-configurable ndd variables preferably are added to the hme_aram_array[] structure in the hme.h file:

trunk-id
trunk-head
flow-type

- 50 - this parameter indicates whether destination MAC address or TCP/UDP destination port-id should be used for flow identification: e.g., 0 denotes use destination MAC address and 1 denotes use TCP/UDP port id;

In the "hme" structure for each interface, add:

```

    unsigned int hme_trunkid;
    unsigned int hme_trunkhead_flag; /* indicates whether
5      this

10     device is the trunkhead */
11     struct hme_trunk*hme_trunkp; /* pointer to an array of
12       hme_trunk structures for the trunk with trunkhead as
13       the first entry */
14
15     unsigned int hme_linkmap_refcnt; /* no. of entries in the
16       linkmap table pointing to this trunk-member */
17
18     /*New structure definitions */
19
20     /*trunk related information which is dynamically created
21       when the trunkhead is identified */
22
23     struct hme_trunk {
24
25       kmutex_t tr_lock; /* lock to protect this
26       structure */
27       unsigned int tr_flowid_type; /*whether
28       dst mac id or tcp/ udpport id*/
29
30       struct hme_flowid tr_last_flowid; /* flowid of the
31       last packet */
32
33       unsigned int tr_last_link_id; /*index of the link used
34       for last packet */
35
36       struct hmep*tr_members; /*pointer to an array of
37       trunk members. first member is the trunkhead */
38
39       unsigned int tr_trunksize; /*no. of trunkmembers in
40       the above array */
41
42       struct hme_linkmap *tr_linkmap; /* pointer to the
43       linkmap table */
44
45       unsigned int tr_linkmap_size; /* no. of entries in the
46       linkmap table */
47
48     } ;

```

Functions for Trunking:

hmeioctl

- 5 - adds code to handle the new ndd variables trunk-id, trunk-head and flow-id;
 - when a new trunk-head is specified, allocates a structure for the trunk and initializes the trunk-member structures and initialize trunk-member hardware devices by calling hmeinit();
 - adds code support ndd variables that contain the instance number, i.e. variables of the form <nn>:<name> where nn is the instance number and name refers to the parameter.

10 hmedetach

- if trunked, deallocates all dynamic data structures allocated for trunk support;

15 hmeareq

- adds code to return error if the device is configured as a non-trunk-head trunk-member;

hme_trunk_init

- 20 - new function: wherever hmeinit() is called to change MAC parameters/mode, calls hme_trunk_init() so that the MAC address, multicast addresses, promiscuous-mode etc. are set for all trunk-members;

25 hme trunk_start

- new function: wherever hmestart() is currently called to transmit a packet, calls hme trunk_start() to handle packet transmission for a trunk;
- this function will use flow-id for the packet to determine the link to be used; if the packet can be posted to a link, it will call hmestart() to transmit the packet on that link; if the packet cannot be posted on any link, it will be put back in its STREAMS queue and the tr_txall_wait flag is set;

hmeintr

- 35 - if link status indicates link-down condition for a trunk-member, assigns an alternate trunk-member, and starts a time-out before using the new-trunk member;

hmeread

- 40 - after getting the received packet in an mblk, use the "hmep" pointer for the trunk-head to send up the mblk;

Modifications and variations may be made to the disclosed embodiments without departing from the scope of the invention.

45

Claims

1. In a network system, a method of dynamically controlling data flow between physical links coupleable to said system, the method comprising the step of:

50 logically combining multiple said physical links into a single logical channel trunk.

2. The method of claim 1, wherein:

55 each of said physical links comprising said trunk is representable by identical physical layer and identical media access control layer characteristics.

3. The method of claim 1, further comprising:

dynamically controlling said data flow to substantially balance data flow carried by each of said physical links comprising said single logical channel trunk.

4. The method of claim 1, further comprising:

5 a server in said system assigning a single media access control layer address to said single trunked logical channel.

- 10 5. The method of claim 4, further comprising:

constructing said single media access control layer address using map-table information available to said system.

- 15 6. The method of claim 4, further comprising:

randomly selecting a destination address for one of said physical links.

- 20 7. The method of claim 6, wherein randomly selecting includes hashing destination media access control layer addresses for said physical links.

8. The method of claim 1, further comprising:

distributing data packets from a server in said system to chosen ones of said physical links such that temporal order of data flow is preserved.

- 25 9. The method of claim 8, further comprising:

said server distributing said data packets over a common segment of said single logical channel trunk.

- 30 10. The method of claim 1, further comprising:

defining, for a server in said system representable by a multi-layer model including at least a physical layer and a network layer, a pseudo-driver software layer disposed intermediate said physical layer and said network layer,

- 35 11. The method of claim 10, wherein:

said pseudo-driver software layer functions as a multiplexer in a receive path and functions as a de-multiplexer in a transmit path.

- 40 12. The method of claim 1, wherein said network is Ethernet-compatible.

13. The method of claim 1, wherein said network functions in full-duplex mode.

- 45 14. A network system, comprising:

a first physical link, coupleable to said system;

a second physical link, coupleable to said system;

50 a server having a server interface that dynamically combines each of said physical links into a single logical channel trunk.

15. The network system of claim 14, wherein:

each of said physical links comprising said trunk is representable by identical physical layer and identical media access control layer characteristics; and

55 said server interface assigns a single media access control layer address to said single trunked logical channel.

16. The network system of claim 14, wherein said address is generated with an address generator selected from a

group consisting of (i) a map-table, (ii) a random selector of destination media access control layer addresses for said physical links, and (iii) a hash unit that hashes destination medial access control layer addresses for said physical links.

5 **17.** The network system of claim 14, wherein said server interface dynamically balances data flow carried by each of said physical links comprising said single logical channel trunk.

10 **18.** The network system of claim 14, wherein said server distributes data packets to chosen ones of said physical links such that temporal order of data flow is preserved.

19. The network system of claim 14, wherein:

15 said interface includes a pseudo-driver software layer disposed intermediate a physical layer and a network layer in said server;

16 said pseudo-driver software layer functioning as a multiplexer in a receive path and functioning as a de-multiplexer in a transmit path.

20 **20.** The network system of claim 14, wherein said network system has at least one characteristic selected from a group consisting of (i) said network functions in full-duplex mode, (ii) said network is an Ethernet network, and (iii) said network is Ethernet-compatible.

25

30

35

40

45

50

55

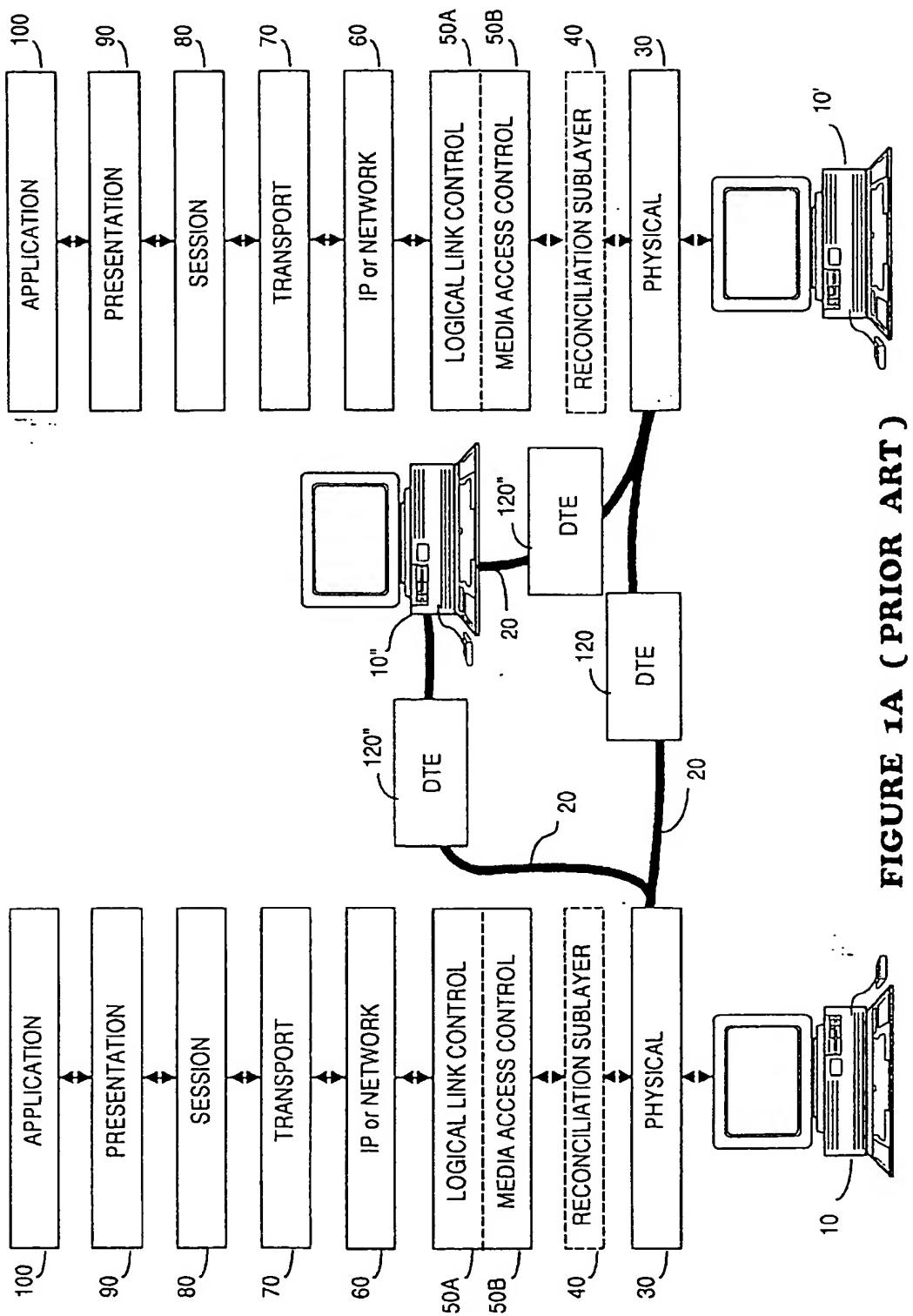


FIGURE 1A (PRIOR ART)

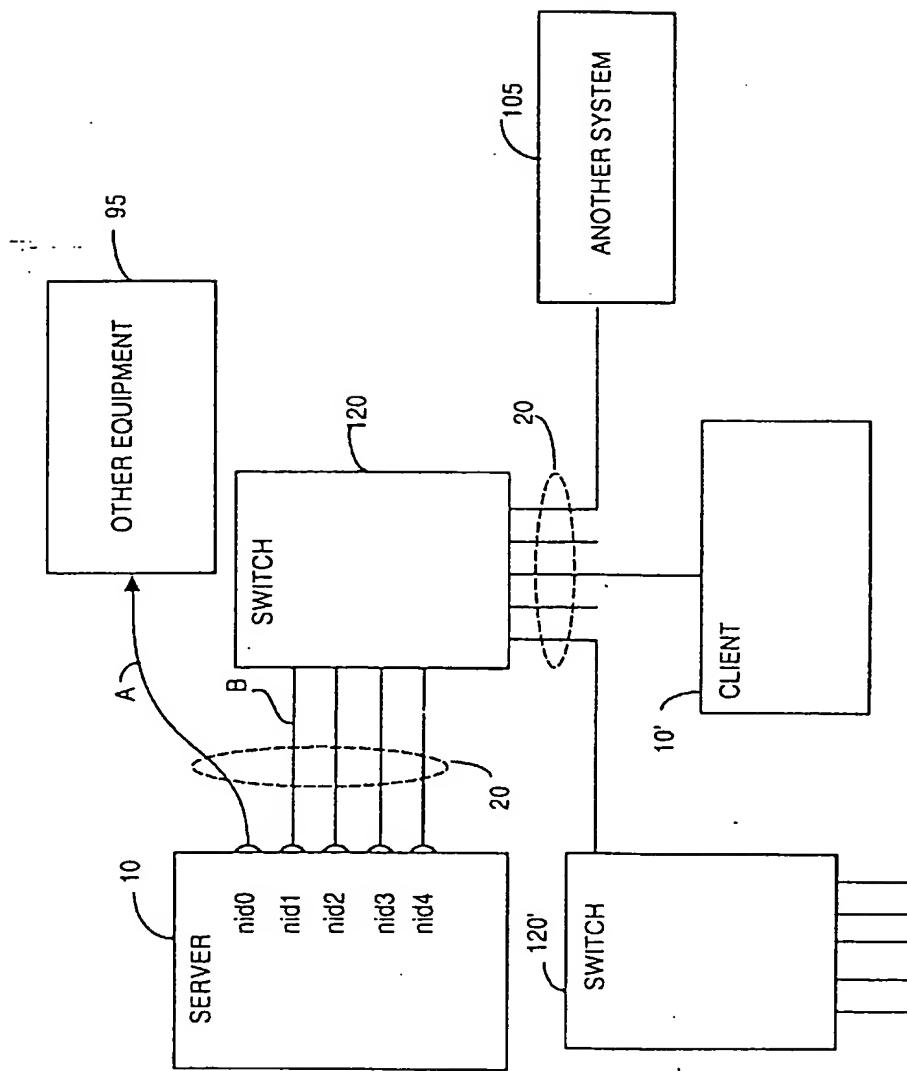


FIGURE 1B (PRIOR ART)

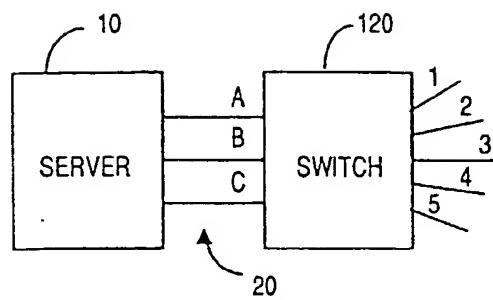


FIGURE 2A
(PRIOR ART)

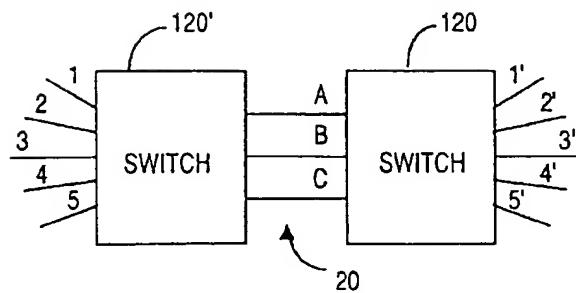


FIGURE 2B
(PRIOR ART)

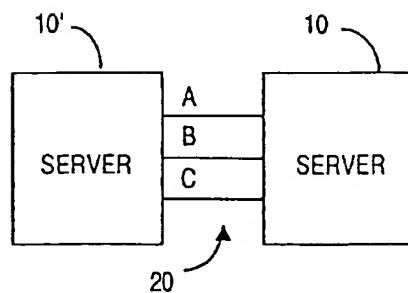


FIGURE 2C
(PRIOR ART)

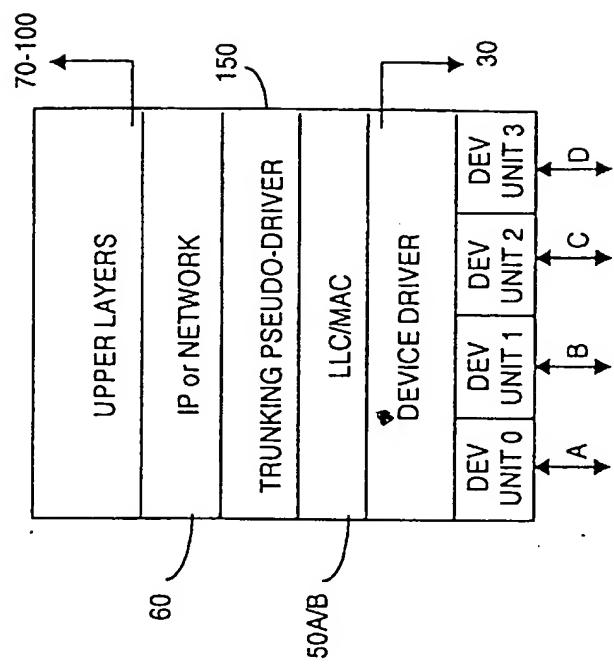


FIGURE 3

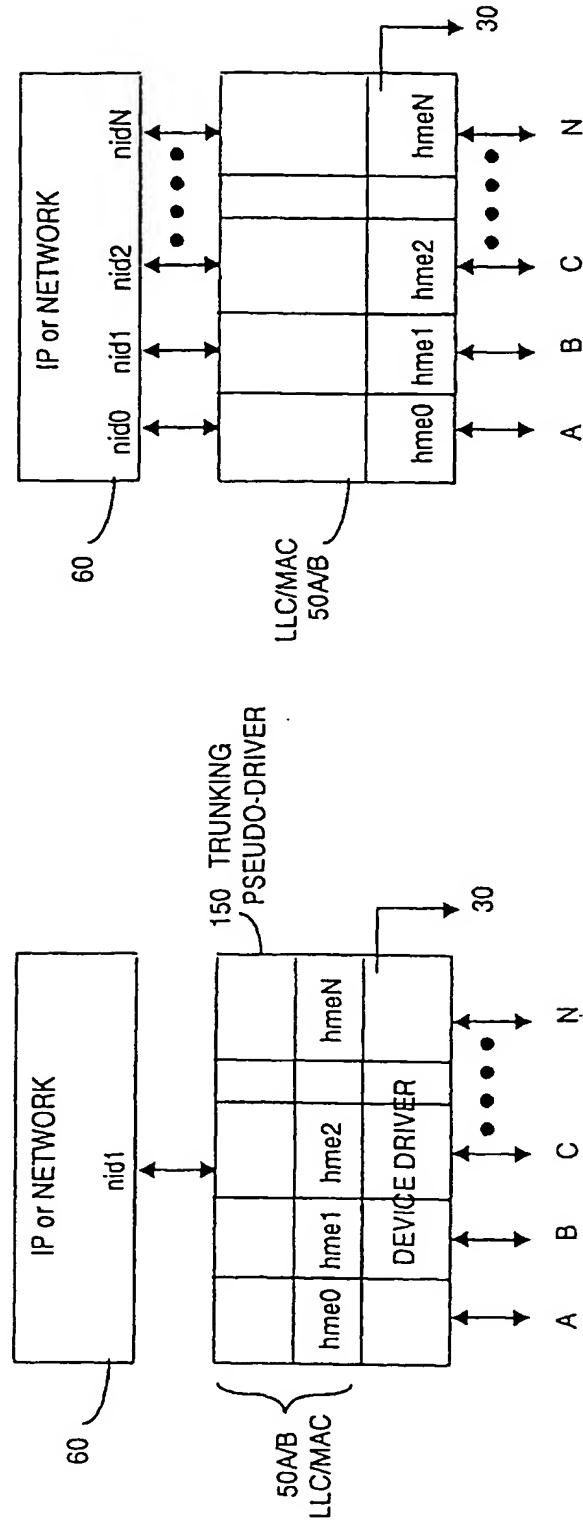


FIGURE 4A

FIGURE 4B
(PRIOR ART)



European Patent
Office

EUROPEAN SEARCH REPORT

Application Number
EP 98 30 5137

DOCUMENTS CONSIDERED TO BE RELEVANT			CLASSIFICATION OF THE APPLICATION (Int.Cl.)						
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim							
X	US 5 633 810 A (MANDAL KALLOL ET AL) 27 May 1997 * figures 4,5 * * column 2, line 21 - column 3, line 16 * * column 4, line 35 - column 5, line 44 * ---	1-3, 8-14, 17, 18, 20	H04L29/06 H04L12/56						
X	US 5 517 488 A (MIYAZAKI SATOSHI ET AL) 14 May 1996 * figures 1-12 * * column 1, line 1 - column 11, line 7 * ---	1-3, 12, 14, 17, 20							
A		4-11, 15, 16, 18, 19							
X	"LOAD BALANCING FOR MULTIPLE INTERFACES FOR TRANSMISSION CONTROL PROTOCOL/INTERNET PROTOCOL FOR VM/MVS" IBM TECHNICAL DISCLOSURE BULLETIN, vol. 38, no. 9, 1 September 1995, pages 7-9, XP000540166 * the whole document * ---	1-3, 10, 11, 14, 17, 19							
A	NISHIZONO T; YOSHIDA Y: "ANALYSIS ON A MULTILINK PACKET TRANSMISSION SYSTEM" ELECTRON. COMMUN. JPN. 1, COMMUN. (USA), vol. 68, no. 9, September 1985, pages 98-104, XP002081618 NEW YORK (US) * abstract *	6, 8, 16, 18	TECHNICAL FIELDS SEARCHED (Int.Cl.) H04L G06F						
<p>The present search report has been drawn up for all claims</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 33%;">Place of search</td> <td style="width: 33%;">Date of completion of the search</td> <td style="width: 34%;">Examiner</td> </tr> <tr> <td>THE HAGUE</td> <td>27 October 1998</td> <td>Eraso Helguera, J</td> </tr> </table> <p>CATEGORY OF CITED DOCUMENTS</p> <p>X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document</p> <p>T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document</p>				Place of search	Date of completion of the search	Examiner	THE HAGUE	27 October 1998	Eraso Helguera, J
Place of search	Date of completion of the search	Examiner							
THE HAGUE	27 October 1998	Eraso Helguera, J							

THIS PAGE BLANK (USPTO)